DOCUMENT RESUME

ED 076 705 TM 002 714

AUTHOR Emrick, John A.
TITLE An Experimental Evaluation of New Measures of Cognitive and Non-cognitive Performance for Elementary School Children.
PUB DATE 73
NOTE 66p.; Paper presented at annual meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Cognitive Tests; Evaluation; Kindergarten; Motivation; Primary Grades; Problem Solving; Self Concept Tests; Self Esteem; Tables (Data); Technical Reports; *Test Reliability; *Test Selection; *Test Validity; Verbal Communication

ABSTRACT

Two experiments to develop psychometric and administrative data on instruments designated for a testing program with disadvantaged elementary school children were undertaken. These instruments provide measures of growth and development in such diverse areas as verbal expressiveness (the ITPA and the Hertzig/Birch scoring of the PSI), problem solving (Raven's Progressive Matrices), self-esteem (Faces and Coopersmith), and achievement motivation (Gumpgookies and Locus of Control). Both experiments involved test-retest assessments of reliability, and factorially balanced assessments of tester effects. The results are discussed with regard to reliability, validity, and suitability of these instruments within and across grade levels K-3. (Author)

FORM 6510
PRINTED IN U.S.A.

FILMED FROM BEST AVAILABLE COPY

ED 076705

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

AN EXPERIMENTAL EVALUATION OF NEW MEASURES OF COGNITIVE AND NON-COGNITIVE PERFORMANCE FOR ELEMENTARY SCHOOL CHILDREN

by

John A. Emrick

Stanford Research Institute

TM 002 214

American Educational Research Association

Annual Convention

New Orleans, 1973

AN EXPERIMENTAL EVALUATION OF NEW MEASURES OF COGNITIVE AND NON-COGNITIVE PERFORMANCE FOR ELEMENTARY SCHOOL CHILDREN

by

John A. Emrick

Stanford Research Institute

This report is based on the results of two experiments undertaken to determine the relative psychometric and administrative merits of a variety of cognitive and non-cognitive instruments being considered for inclusion in the National Follow Through Evaluation Battery (Emrick et al., 1973). At the time these studies were undertaken, the FT battery consisted primarily of tests of conventional academic achievement (reading, language skills, mathematics) for grades K through three (SRI, 1969, 1970). Only a limited number of measures had been directed toward assessing children's social and emotional development, and the results of a previous attempt to develop and incorporate alternative measures of non-cognitive growth were, at best, equivocal (SRI, 1970). To help meet the need to strengthen the FT battery both in scope and depth, a number of additional instruments and measuring techniques were identified and experiments were designed and conducted to develop relevant psychometric and administrative data on each such instrument. These data were to serve as a basis for subsequent inclusion of tests in the FT battery, for administrative planning, and for interpretation of results.

These instruments and measuring procedures can be grouped into four categories in terms of the psychological traits purportedly assessed. These categories, and associated instruments, are as follows:

- Affect (primarily self-esteem)

  The "Faces" Attitude Inventory (SRI, 1971)

  The Brown IDS Self Referent (Brown, 1971)

  The Coopersmith Self-Esteem Inventory (Coopersmith, 1970)

- Achievement Motivation

  The ETS Locus of Control Test (ETS, 1970)

  The Gumpgookies Test (Adkins, 1970)

- Verbal Expressiveness

  The "Verbal Expression" subtest of the Illinois Test of Psycholinguistic Abilities (Kirk et al.1969)

  The Hertzig/Birch scoring procedure for the Preschool Inventory (Hertzig et al., 1968)

- Problem Solving

  The Coloured Progressive Matrices (Raven, 1962)

Although the above grouping is somewhat arbitrary, it does reflect some correspondence between the test items and the psychological theory underlying the tests.

Measures of Affect. Measures of affect are commonly regarded as important to an adequate understanding of the emergence of social and behavioral skills. Instruments in this group measure how the child feels about learning, about himself, and about others; also included are measures of how the child reports he feels that others (teachers, parents and peers) perceive and regard him. Interest in measures of

affect generally derive from either of two competing theoretical positions. One the one hand, attitudes are argued to predispose behavior, such that as the individual develops more positive feelings regarding himself and others, he will be more likely to develop the psychological and social requisites for advancement and self-sufficiency. On the other hand, positive affect can be considered a consequence of both success experiences and of more favorable peer and adult-child interpersonal relations, hence primarily reflecting the results of concomitants of growth (rather than the antecedents). In terms of intervention programs such as Follow Through, the first interpretation suggests that as the child experiences more academic success (which he should in the Follow Through program), he will develop a more positive attitude about himself and about learning, and, correspondingly, he will enjoy improved interpersonal relationships with other children. The second implies that as the child is given more positive evidence of his own personal worth through his relationships with teachers and adults who treat him with positive regard, he will develop more feelings of self-worth and more positive attitudes toward learning and toward others.

The three instruments used to assess child attitudes and feelings were the "Faces" Attitude Inventory, a modification of the Brown IDS Self Referent Test, and the Coopersmith Self-Esteem Inventory. All of these instruments involve a self-report measurement procedure in which a question is presented to the child and he then chooses from among the alternatives the one that best describes his feeling. The Faces test

requires that the child choose from each of 3 alternative faces (happy to sad, or neutral) to answer each of 16 items. The first two items are practice, and the remaining 14 ask the child to show how he feels (choose a face) regarding school, learning, peers, self, and how parents, teachers, and peers feel toward him. For example, Item 1 reads: "Think about having fun. Point to the face that shows how you feel when you are having fun." This "choose a face" response device is considered desirable since it is (a) relatively free of linguistic requirements and (b) enables very young children a relatively unambiguous means of responding to the items. Actually, two forms of this response device were employed in one of the experiments: a smile-to-neutral and a smile-to-frown version.

The Brown IDS Self Concept Referents (Brown, 1966) Test was also originally developed with an eye toward reducing response ambiguity and enhancing the reliability of resultant data with young children. This instrument, based on G.H. Mead's (1956) model of the development of self-awareness, requires the child to describe himself in the third person by selecting (forced choice) from each of 14 bipolar adjectives or descriptions. For example, Item 1 reads "Is (pointing to picture and saying child's name) happy or is (child's name) sad?" Of the four original referents (self, mother, teacher, peer), the modification used in this study employed only the self (polaroid picture of the child) and the teacher. Also, only 5 of the 14 bipolar descriptions were used with the teacher referent. These modifications were imposed primarily to reduce the administration time required for the test.

The Coopersmith Self-Esteem Inventory--originally developed for children of 8 years and older (Coopersmith, 1965) was slightly modified for this study. This test requires the child to select from two alternatives to each of 55 descriptive statements. For example, Item 1 reads "I spend a lot of time day dreaming," to which the child responds "like me" or "not like me." This instrument includes 4 affect scales (self, peer, home, school) and a lie scale, although our treatment of data aggregated responses into a single affect--or self esteem--scale (exclusive of lie-scale responses).

Measures of Achievement Motivation. The second group of measures have to do with how the child conceptualizes and values his academic and/or social success or failure. Achievement motivation usually refers to an individual's tendency to approach tasks of varying difficulty and to stay at these tasks until some degree of success is achieved. Research in this area (Maehr and Sjogren, 1971) generally has shown that people characterized as high on achievement motivation tend to be attracted to moderately difficult tasks where the probability of success vs failure is about equal. People low in achievement motivation, on the other hand, tend to either engage in very difficult tasks--and therefore cannot be blamed for failing--or in extremely easy tasks in which the probability of failure is almost zero.

The assumption appears to be that if an intervention program like Follow Through is successful, it will tend to generate more positive achievement motivation in pupils as they proceed through the program.

This will be due in part to the children's new experience with success in learning activities that bring learning into the "moderately difficult" range. The children will also begin to learn that success and failure are under their control. The Locus of Control test--i.e., the way and degree to which a child attributes successes and failures to either internal (within himself) or external (events beyond his control) causes--used in this battery offers two alternatives per item from which the child selects the one he feels best describes his situation. Each of these options can be characterized as reflecting either internal or external locus of control. It is also possible to differentiate between children in terms of their locus of control on successes vs failures. That is, some children may feel responsible for their successes but feel that something else is responsible for their failures, and vice-versa.

Four versions of this instrument were originally developed (ETS, 1968) to control for possible sex and ethnic biases. In developing a form that could be group administered, and in the absence of data supporting the presumed ethnic biases, the four versions were consolidated into a single form with items balanced for sex and ethnic characteristics. Ten additional items were added to the basic measure which repeat a previous item but reflect a change in sex (5 items) or ethnic characteristics (5 items). The latter items allow sex and ethnic effects to be investigated if they are present. For present purposes, only the twenty-two items directly comparable to the ETS version will be analyzed.

The alternate achievement motivation instrument is the Gumpgookies test (Adkins & Bailiff, 1970, 1972). This name was selected by the author of the test primarily to make it intrinsically interesting to the very young children for whom the test was designed initially. Gumpgookies was included in this study to evaluate its appropriateness with older (2nd and 3rd Grade) children.

The Gumpgookies is similar both to the Locus of Control and, to some extent, to the Coopersmith Self-Esteem Inventory. The logic behind the Gumpgookies test is that the child can, through use of his imagination and, to some extent, his projection, imagine a Gumpgookie of his very own which does everything he would do. In this fashion the child is able to choose among a number of dichotomous alternatives by letting his Gumpgookie (his alter-ego) do the choosing for him. Recent research (Adkins & Bailiff, 1972) has identified five factors--or dimensions of achievement motivation as measured by this instrument. These are:

1. Instrumental activity (thinking of and doing those appropriate activities that are instrumental to achievement)
2. School enjoyment (positive attitude toward school)
3. Evaluative (awareness of one's abilities and excellence)
4. Self confidence (being best, coming out on top)
5. Purposive (awareness of implications of present behavior for the future)

This factor analytic interpretation of the Gumpgookies instrument is still undergoing refinement. Consequently, data reported in this study treat item responses only on the more general achievement motivation level.

Problem Solving Measure. Only one problem solving measure is included in this study. The problem solving test used in this investigation is the Raven's Coloured Progressive Matrices (Raven, 19). Each item consists of a pattern and a series of choices for a component missing from the pattern. The child has to study the pattern, determine the logical components, and then identify from the choices the component that completes the pattern. Although this test was originally designed as a "culture fair," non-verbal intelligence test, it is used here as a problem solving test.

Verbal Expression Measures. The fourth area of measurement included in this study was that of verbal expressiveness. Since much emphasis is placed on the early acquisition of language skills, and since both cultural and psychological explanations of performance deficits among disadvantaged populations rely heavily on linguistic factors, measures of psycholinguistic development were included in this study. The two instruments studied were the Hertzig Birch scoring procedure for the Preschool Inventory and the Verbal Expression ubtest of the ITPA.

The Preschool Inventory Test (PSI) is designed to sample the child's general information, his familiarity with a number of common objects, events, and attributes, e.g., How many wheels does a bicycle have? When do we eat breakfast? The Hertzig Birch scoring convention of the Special Edition allows for an expanded coding of the "correctness" or "incorrectness" of pupil response to each item and for indicating the occurrence of verbal behavior. There are two correctness codes and six incorrectness codes.

The two correctness categories are:

A simple, correct response.

A correct response where the child elaborates on his own (spontaneously).

The six incorrectness categories are:

A simple, wrong response.

A wrong response in that the child actually refuses to answer the question.

A wrong response in that the child actively pursues some substitute activity.

The child says, "I don't know" or some equivalent.

A request for aid. That is, he asks the tester to work the problem or find the answer for him (which is not the same as the child asking the tester to clarify the question.

The child makes no response. He simple remains mute.

This scoring procedures allows for three categorizations of responses to the test items:

(a) Competency--or the number correct

(b) Verbalization--or the extent of spontaenous verbal behavior during testing

(c) Style--or the tendency to attempt answers regardless of correctness of responses.

The Verbal Expression Subtest of the Illinois Test of Psycholinguistic Abilities (ITPA) is designed to provide a measure of the child's verbal expressiveness. Moreover, based on developmental work (Paraskevopoulos & Kirk, 1969), normative interpretations in terms of psycholinguistic age are available, provided the test is properly administered. The administration of the Verbal Expression Subtest involves handing the child a familiar object, asking him to describe the

object, and writing down his response. An aluminum nail 2 inches long; a small red rubber ball; a green wooden block with 1-inch sides; a plain white envelope; and a white plastic button 1½ inches in diameter are shown to the child in that order. The first item, a practice item, is used to familiarize the child with the test procedure and to suggest to the child the types of verbal expressiveness he can engage in. When administering this test, the tester must know when and how to probe for clarifications and how to record the child's responses. These tasks are very complex due to the flexibility of the language, and the principal concern in investigating this instrument was the extent to which such administrative and interpretative skills could be developed reliably among paraprofessional testers.

EXPERIMENT 1

This experiment was designed to establish the operating characteristics (mean, variance, reliability, standard deviation) and tester bias on four instruments designed for inclusion in the national evaluation battery, entering level (kindergarten, or first grade in schools without kindergarten). One purpose was to determine which, if any, of the instruments under consideration produced the most reliable and efficient measures of affect (Faces vs Brown) and of verbal expression (PSI-Hertzig Birch vs ITPA). Another purpose was to establish baseline performance on a national cross-section of pupils for eventual follow-up assessment to determine the predictive validity of each of these instruments.

The overall purpose of this experiment was to develop information to help guide future operational decisions in the evaluation of Follow Through. The specific immediate questions posed by this experiment were:

(a) What are the psychometric properties, administrative requirements, and relative costs and economies associated with each of the four instruments?

(b) What are the relative convergent and discriminant validities of these four tests as administered to a common population?

(c) What are the relative sensitivities of each instrument to tester differences, etc.?

Table 1

SUMMARY OF TEST STATISTICS

FOR EXPERIMENT 1

	PSI(H/B)	BROWN	FACES	ITPA
Mean				
Test	16.7	16.6	5.9	13.8
Retest	18.1	17.5	6.6	13.8
Standard Deviation				
Test	5.71	4.00	2.56	6.55
Retest	5.64	3.54	3.00	6.12
Reliability				
Test	.834	.816	.557	.820
Retest	.839	.787	.697	.825
Standard Error				
Test	2.15	1.55	1.63	2.78
Retest	2.10	1.42	1.60	2.56
Number of Cases				
Test	651	692	669	675
Retest	654	671	648	657

METHOD

Subjects. Altogether 17 separate schools located throughout the contingent U.S. were included in the experiment. In 16 of 17 locations two classrooms at the entering grade level (primarily kindergarten) participated in the study; at the remaining location, only one classroom participated. All of these classrooms were also participating in the National Follow Through Evaluation. The locations were selected so as to be minimally representative of the overall distribution of Follow Through programs in terms of proportion urban, ethnic mix, kindergarten to first grade entrance, and gross program type. Hence, this overall sample, as well as the varieties under which tests are administered are considered reasonably representative to support overall generalizations developed from analyses of data.

Experimental Design. A 2 x 2 factorial design was employed in which test occasion (test or retest) defined one factor and tester (same tester on both test and retest, or different tester on test and retest) defined the other factor. The interval between the first and second test varied from two to three weeks.

In each classroom, half of the students were tested by the same tester on both occasions; half were tested by different testers. Testing was counterbalanced in such a way that (1) one-quarter of the students were tested by Tester A and retested by Tester B, (2) one-quarter were tested by Tester B and retested by Tester A, (3) one-quarter were tested both times by Tester A, and (4) one-quarter were tested both times by Tester B.

The discrepancy between the number of students assigned to a particular test-retest condition and the number of students for whom both test and retest scores were actually obtained is apparent in the following tabulation:

TEST

Data Category	Tester Condition	PSI	ITPA	Brown	Faces
Number of cases assigned to test-retest conditions	Same	336	336	331	331
	Different	339	339	338	338
	Total	675	675	669	669
Number of cases having both initial and retest data	Same	274	286	295	332
	Different	315	327	224	263
	Total	589	613	519	595

The 15% shrinkage evident in this tabulation occurred due to normal attrition, absenteeism, and tester problems typically encountered in field-experimental studies.

Procedure. Several field supervisors from the SRI staff, most of whom had also assisted in the earlier field trials, prepared themselves to conduct special training for other testers in the field. Members of the analysis staff also assisted in this training.

During supplemental testing, tests were administered at each location to pupils by a team consisting of a supervising tester and three assistants. The supervising tester and one test assistant administered the PSI and the ITPA; the other two test assistants gave the Brown and the Faces tests. All tests were administered individually.

Approximately three weeks later, all tests were re-administered to the same pupils by the testing staff in accordance with the experimental design. Variations in retest interval were due to scheduling problems at different locations. However, this variation is small and presumed negligible in impact.

RESULTS AND DISCUSSION FOR EXPERIMENT 1

Two sets of analyses were performed on each of the instruments. The first set of analyses involved the computation of summary and item test-retest statistics. The second set of analyses is comparative across tests and includes the computation of inter-test correlations, retest reliability, analysis of variance of tester effects, and evaluation of administrative problems.

PART I: ITEM AND TEST STATISTICS

The Pre-School Inventory (PSI), Hertzig Birch Scoring

A summary of test-retest statistics for the total sample is presented in Table 1. Included in this summary are test and retest values for the mean, standard deviation, reliability (Kuder-Richardson formula 20), standard error of measurement, and number of children tested with the PSI. Inspection of this table reveals the overall mean score on the initial administration of the PSI was 16.7 correct responses, or about 58% correct, with a standard deviation of 5.71. Initial test means varied considerably across projects from a low of 12.6 to a high of 20.0.

A systematic score increment from initial test to retest is evident in this table. An overall improvement of one and one-half correct responses occurred on the retest, but the standard deviation of the test remained essentially constant. An upward score shift occurs without exception in each project.

The reliability of this test, as estimated by the KR 20 formula, remains markedly stable from test (.834) to retest (.839). However, these estimates suggest the test is only moderately discriminatory with regard to individual differences. That is, these reliability values indicate a standard error of measurement of slightly more than 2 raw score points. Thus, 99% true score confidence interval for child scores is approximately 8 raw score points (± 4 points). And surprisingly, although there is considerable variability across projects in the estimated reliability for both initial testing (.673 to .946) and retesting (.562 to .933), variation in standard errors remains small (1.96 to 2.31 on initial testing, 1.92 to 2.34 on retest).

When scores aggregated up to a classroom level, the reliability of this test (now in terms of detecting differences between classrooms) substantially increases. For example, the estimated reliability becomes in the neighborhood of .99 and the standard error of measurement reduces to approximately .4. Thus, classrooms which differ by 2 or more raw score points differ significantly.

A somewhat more detailed item analysis of this test reveals a large variation in average item difficulties both across items and across projects in the sample.

However, the overall difficulty levels of the items appear generally appropriate for this sample of children in that most items are within a 30%-70% difficulty level range.

A further analysis of individual item responses shows a surprisingly high proportion of "attempts" regardless of the correctness of the response (i.e., the work vs competency distinction). For both pre- and posttests, more than 90% of the items were attempted overall (pre = .93, post = .94), of which more than 2/3 were correct. Also, verbalizations occurred in more than 1/2 of all responses.

In summary, the evidence in terms of the distributional statistics, operating characteristics, and item analyses of the PSI--using the Hertzig Birch scoring procedure--is positive. The test in this form has satisfactory reliability in terms of group discriminability, appears appropriate to the age range, exhibits a desirable and stable distribution of item difficulties, and appears suitably adaptable to the Hertzig Birch scoring technique.

The Brown IDS Self-Referent

The psychometric data for the Brown test are summarized in Table 1. Data indicate the sum of positive affect (A+) responses occurring for the 21 items. These summary statistics reveal a retest increment of nearly 1 A+ response, reasonably high reliability (KR90) and very low standard errors of measurement. But inspection of data across location revealed several instances of probable ceiling problems (up to 93% A+).

The overall initial test reliability is .816 and the retest is .787. If the location with the ceiling problem is omitted (its reliability estimate is .143), the retest reliability estimate is .827. Further,

these values are remarkedly uniform across locations, suggesting the test is relatively unaffected by subgroup characteristics.

Inspection of item analysis data suggests this improved discriminability was primarily due to reduction of ambiguous responses on the retest: the subjects appeared better able to respond to the items. However, the high homogeneity of item responses suggests the instrument is most probably measuring a unidimensional trait.

In sum, the analyses of test-retest data for the Brown IDS Self-Referent Test have indicated this test has enigmatic properties. On the one hand, it appears to have respectable reliability and low measurement error, particularly for tests of this domain (non-cognitive). On the other hand, scores obviously suffer from ceiling problems and a careful study of item properties suggests the apparent high reliability is possibly due to item artifacts. Also, a hint of social expectancy response bias is evident from item analyses.

The "Faces" Attitude Inventory

The test-retest statistics for the "Faces" inventory of pupil attitudes about self, school, and learning are summarized in Table 1. The scores interpret as the "positive atttitude" or number of A+ responses.

The mean A+ score on the initial test was 5.9 or 42% (SD = 2.56). This mean varied across projects from a low of 4.8 (34%, SD = 2.81) to a high of 7.8 (56%, SD = 2.53). In all, the interproject variation of mean A+ scores was relatively small, and the variances were surprisingly uniform.

On retesting the A+ means and variance become less uniform across projects, and an increase over the initial score is apparent. The instrument appears more reliable on retest than on initial testing (.557 and .697 respectively). This increase is apparently due to the increased score variance for restesting which occurred in three projects, and the corresponding increases in estimated reliability. In one instance (Project 3002), a shift in test-retest standard deviations of 1.91 to 3.00 corresponded with reliability estimates of .191 and .713, respectively. One implication of this finding is that for certain projects substantial improvements in the testing procedure (administration, testwiseness, or both) occurred.

Comparison of test-retest standard errors shows that the 99% confidence interval for this test is approximately six raw score points ($\pm$ 3). At the classroom level, the interval reduces to approximately 1.25 points. Thus, classrooms whose average A+ scores differ by more than $1\frac{1}{2}$ points can be interpreted as significantly different on this measure.

Item analyses show that for the initial test, the items tend to yield strikingly uniform overall proportions of A+ choices. This item homogeneity becomes even more pronounced on the retest, and all A+ proportions tend to shift upward. The patterns of responses by response option indicate that majority were to the A+ choice, and the remainder were divided somewhat evenly on the A_0 and the A- choices.

To suggest the homogeneity of the item response profiles is an artifact of some random response set or other bias is not supported by the data. First, a random set would suggest 33% as a central tendency,

where the data are closer to 50%. Second, huge inter-project variance in percent A+ responding by item is observed in terms of the project by item patterns. Consequently, whatever misgivings one may have regarding the face validity of this measuring instrument, it does display interesting and somewhat desirable psychometric properties.

In summary, the evidence in terms of the distributional statistics, operating characteristics, and item analyses of the "Faces" attitude inventory used in this study is modestly positive. Although the reliability of this test (KR 20) was not particularly high, it appears comparable (if not above) that obtained with similar instruments. Particularly uniform and stable estimates of measurement error suggest the instrument has relatively robust properties and may be useful as a method of detecting group differences. Also, substantial inter-item homogeneity was evident. Inter-project item response profiles are considerably more heterogeneous, suggesting the instrument may be sensitive to inter-project differences, and potentially useful in detecting same.

The Illinois Test of Psycholinguistic Abilities: Verbal Expression Subtest (ITPA)

The SRI adaptation of the verbal expression subtest of the ITPA presented perhaps the most complex administration and processing problems of all instruments included in this study. The difficulty in analyzing results for this test lies in the complexity of scoring responses. In an attempt to remain faithful to the guidelines and scoring instructions contained in the author's manual, recorded test responses were converted

to the number of discrete and scoreable descriptions for each item and then tabulated as item scores. These item scores were then summed to yield a total expressiveness score for each child on each test occasion.

These data were then analyzed to display the psychometric properties of the test, and as summarized in Table 1, show virtually no test-retest effect on mean performance. This finding is contrary to the clear pattern of retest increments observed for the previous three instruments. Also, the mean total response score of the four items closely replicates the normative data reported by test authors (Paraskevopoulos and Kirk, 1969) --i.e., the norm for pupils 5-5½ years is 14 points, whereas the average in this study was 13.8.

Although the internal consistency reliabilities (Cronbach, 1970) are very stable and within "acceptable" range, high test variance has produced relatively large standard error estimates. For example, the estimated standard error of the initial overall test scores is 2.78. The 99% confidence range becomes +5.5 score points or 11 points. Although this result indicates a potential floor problem, this standard error is well within the range for this age level as reported by the authors in their standardization study.

Estimating the standard error of classroom scores as .5, classrooms showing average score differences of more than 2 score points will be significant. However, inter-location variance is large and significant on this test.

The results of item analyses suggest a possible problem with test administration occurred due to inappropriate use of the practice item. Since the testing instructions require that the tester elicit at least five classes of appropriate verbal responses from the child (through use of hints, prompts, and probing), the minimum pretest average should be five. However, most locations showed averages below 5 suggesting the scores may be underestimates of the children's verbal expressiveness.

In sum, the results of these analyses are somewhat ambiguous regarding the appropriateness or utility of wide-scale testing with the ITPA. The overall mean response scores are fairly close to those obtained in more controlled(clinical) testing conditions under more refined testing procedures. Problems in administration may have acted as a score suppressant. The reliability and measurement error of the scores appear acceptable and comparable to standardized data but test variance and inter-project variance would make interpretation difficult.

However, since item response profiles for the present data are strikingly similar to those reported in the technical manual, concurrent validity is implied.

PART II: COMPARATIVE ANALYSIS OF TESTS

To evaluate for relative test-retest stability of scores as well as for the relative common and unique properties of different tests, a test-retest intercorrelation matrix was constructed. This matrix, as presented in Table 2, has been organized to show three correlational properties of the tests as obtained from the experiment:

(a) the retest reliability--or stability of scores over the three-week interval

(b) the concurrent validities--or the relative score overlaps for tests purportedly measuring the same underlying traits, and

(c) the discriminant validities--or the relative orthogonality of tests purportedly measuring different traits.

In Table 2, reliabilities are shown in the solid blocked areas, concurrent validities in the shaded blocked areas, and discriminant validities in the dotted line blocked areas. This organization of intercorrelations allows convenient detection of simplexes regarding reliability, convergent and discriminant validities. For example, a simplex is evident for the PSI and the ITPA tests. Respective retest reliabilities are .85 and .61,* concurrent validities range from .47 to .57, and discriminant validities range from .38 to .15. Hence, these two tests (PSI and ITPA) provide reasonably stable measures of the cognitive behaviors purportedly

---

* More recent data, currently being analyzed, suggest the retest stability of this instrument is substantially higher--more in the neighborhood of .75. Problems in coding and scoring procedures are reflected in the Table 2 estimate.

measured, have a substantial degree of overlap yet are sufficiently independent to suggest different aspects of the traits are represented in the scores, and are essentially independent of the constructs purportedly measured by the Brown and Faces tests.

The noncognitive tests (Brown and Faces), however, do not display a simplex pattern, primarily due to low concurrent validities. Their respective reliabilities are .55 and .65, which is within the range generally obtained for psychological tests of affect. But these scores appear more strongly related to the PSI and ITPA than to each other. This is most likely due to the previously described ceiling problems with the Brown. Actually, the Faces test appears fairly valid in that it has less in common with the ITPA and PSI and has a higher reliability than the Brown (particularly when retest values are inspected). This further supports our interpretation that the data for the Brown reflect--to a large extent--linguistic factors and social desirability response components.

Table 2

TEST-RETEST CORRELATION MATRIX OF THE FOUR TEST INSTRUMENTS

(Sample sizes in parentheses)

	PSI		Brown		Faces		ITPA	
	Test	Retest	Test	Retest	Test	Retest	Test	Retest
PSI								
Test	1.000 (597)							
Retest	.845 (597)	1.000 (597)						
Brown								
Test	.293 (557)	.319 (557)	1.000 (632)					
Retest	.362 (557)	.378 (557)	.545 (632)	1.000 (632)				
Faces								
Test	.315 (505)	.318 (505)	.229 (581)	.171 (531)	1.000 (552)			
Retest	.335 (505)	.334 (505)	.204 (531)	.257 (531)	.646 (552)	1.000 (552)		
ITPA								
Test	.566 (578)	.517 (578)	.248 (582)	.276 (582)	.197 (508)	.204 (508)	1.000 (620)	
Retest	.467 (578)	.517 (578)	.299 (582)	.314 (582)	.150 (508)	.154 (508)	.608 (620)	1.000 (620)

Key

□ = TEST-RETEST RELIABILITY

▨ = CONCURRENT VALIDITIES

⬚ = DISCRIMINANT VALIDITIES

Analysis of Variance of Retest Scores Across Tester Factors

The results of analysis of variance on test scores (same testers vs different testers across the retest interval) are presented in Table 3. These results indicate that significant changes in test-retest scores occur on the PSI ($F = 66.44$, $p < .001$), on the Faces ($F = 72.03$, $p < .001$) and the Brown ($F = 32.71$, $p < .001$), but not on the ITPA ($F = 0.06$). However, in no instance did an interaction of test-retest score by tester condition (same - different) reach significance. This finding is particularly important since it implies, at least to some extent, the independence of these scores from specific tester properties.

There is evidence of an overall tester effect for the Faces test ($F = 5.73$, $p < .05$), the most likely interpretation of which is sampling bias. This is because the initial test score differences remained when testers were factorially crossed for posttesting.

In the lower portion of Table 3 coefficient etas are reported for each test. Again the PSI has the highest estimated value (in this case, non-linear correlation or generic reliability), followed in rank order by the Faces, ITPA, and Brown tests. This rank ordering of values is in exact agreement with that earlier reported for the zero-order Pearson retest correlations.

Table 3

ANALYSIS OF VARIANCE OF TEST-RETEST DATA
BY RETEST CONDITION (SAME VS. DIFFERENT TESTER)

	Measure											
	PSI			ITPA			Faces			Brown		
Source of Variance	df	MS	F	df	MS	F	df	MS	F	df	MS	F
Testers	1	98.97	1.54	1	80.86	1.26	1	81.07	5.73*	1	6.28	0.31
Error (between)	587	64.05		611	64.18		517	14.14		593	20.43	
Retest	1	365.31	66.41**	1	1.06	0.06	1	221.04	72.03**	1	196.85	32.71**
T x R	1	14.34	2.61	1	18.33	1.16	1	0.34	0.11	1	1.95	0.32
Error (within)	587	5.50		611	15.73		517	3.07		593	6.02	

* P<.05
** P<.001

SUMMARY OF MEANS

	PSI		ITPA		Faces		Brown	
	Test	Retest	Test	Retest	Test	Retest	Test	Retest
Same Tester	16.8	18.2	13.5	13.8	6.3	7.2	16.8	17.7
(n)	(274)		(286)		(295)		(332)	
Different Testers	17.6	18.5	14.3	14.0	5.7	6.7	16.8	17.5
(n)	(315)		(327)		(224)		(263)	
Coefficient ETA	.914		.755		.783		.705	

PART II: COMPARISON OF ADMINISTRATIVE PROBLEMS AMONG THE FOUR TESTS

Each tester is required to enter a description of conditions and problems (if any) for each and every test given to every child he (or she) tests. This logging includes entries of testing time as well as explanation of test problems or conditions. To evaluate the relative administrability of each of these tests in terms of the problems commonly encountered in testing situations, a detailed examination of entries in tester logs was performed.

Table 3a presents a summary of the frequency of problems identified by each test by each occasion. As can be seen in this table, the majority of problems occurred with the ITPA, whereas the fewest problems were encountered on the Faces test. Also, fewer problems were encountered in retesting on all tests.

That the ITPA has the largest number of problems, regardless of category or occasion, is not particularly surprising given that this is a very complicated test to administer and score. It is likely that inexperienced testers projected uneasiness to the children, and this in turn produced restless or anxious behavior on the part of the latter. It is also possible that the requirement that responses be recorded verbatim contributed to at least some of the measurement problems for the ITPA.

Table 3a

FREQUENCY OF TESTING PROBLEMS AS INDICATED IN TESTER LOGS

Problem	PSI		Faces		Brown		ITPA		Total
	Test	Retest	Test	Retest	Test	Retest	Test	Retest	
Child had difficulty speaking or understanding English.	32	22	12	5	15	4	41	25	156
Child was restless, inattentive, or difficult to control.	54	42	44	36	47	38	73	52	386
Child would not respond or refused to take the test.	22	22	6	5	6	10	45	31	147
Child was removed from test situation. (Explain)	5	1	5	2	9	3	13	5	43
Child frequently borrowed answers.	5	1	-	-	-	-	6	1	13
Noise distraction.	83	81	61	64	56	66	97	86	594
Disruption by uninvited guest.	20	27	21	5	11	3	23	8	118
Disruption from teacher's presence.	2	2	-	-	1	2	1	-	8
Test materials problem.	-	1	6	1	15	4	1	-	28
This test booklet was not completed. (Explain in detail)	5	2	7	4	5	6	16	6	51
Exclude these data from the analysis. (Explain in detail)	-	-	-	-	-	-	-	-	-
Total	228	201	162	122	165	136	316	214	1,511

SUMMARY

The performance patterns of nearly 700 Entering Kindergarten (and a few Entering First Grade children) were studied on each of 2 testings for each of four tests: The PSI:Hertzig Birch Scoring, the Verbal Expression Subtest of the ITPA, the Brown IDS Self-Referent test, and the Faces Attitude Inventory. These results indicate the PSI displays the most desirable psychometric properties, has the greates: retest stability, and the Hertzig Birch procedure appears adaptable to this test. The ITPA displayed a relatively high standard error, and only modest retest stability. The concurrent validity of the ITPA and PSI are reasonably high, certainly higher than the other tests studied.

The other two instruments--the Brown and Faces--are also purported to measure a single domain; in this instance affective. The analysis of data for the Brown test indicates the instrument has potential ceilin; problems and unacceptable retest reliability. Furthermore, item analysi: suggests test items may not be sufficiently sensitive to examine variabi.ity. The Faces test, on the other hand (and contrary to prior speculation), provided encouraging evidence of its psychometric capability to discriminate and provide stable estimates of group (or individual) differences in reported attitudes. Furthermore, this test appeared less related to measures of the cognitive domain than the Brown, and the Faces and Brown test scores are essentially unrelated (low concurrent validity).

Analysis of variance of test-retest data reveals the characteristic retest score increment as significant for all tests except the ITPA. No tester effects were observed to interact with test-retest score

patterns, and only for the Faces test did any score differences by testers occur (in this instance, the effect was small and unintepretable).

Analysis of tester logs provides additional support for the interpretation that the ITPA may be too complex an instrument to administer in field settings using paraprofessionals. Further studies are currently being conducted to establish the validity of the administration and scoring procedures of the ITPA.

RECOMMENDATIONS

Based on the results of Experiment 1 we advance the following recommendations for testing programs at the kindergarten and first grade level.

(a) That the Hertzig Birch scoring procedure be utilized for any test which is individually administered to young children. This procedure adds little, if any, administration time to testing, can be effectively utilized by any qualified psychological or educational tester, and adds substantially to the richness of data and to the reliability of the measures obtained. With the Hertzig Birch procedure, protocols can be scored for competence (# correct), style (# and type of attempts), and verbal expressiveness.

(b) That the Faces attitude inventory be used where there is a need or desire for a brief test of feelings toward school, self, and others. This test appears appropriate for young children

(K-level) when administered individually.

(c) That the modified version of the Brown, as studied here, not be employed as a measure of self-concept. This test may be appropriate for preschool populations, but evidence of ceiling effects with K and EF samples argue against its adoption.

(d) That the ITPA, Verbal Expression Subtest be used only if a measure of psycholinguistic age is essential. This test is very difficult to administer and score, and further study of our procedure is required to establish the validity of our data.

EXPERIMENT 2

This experiment was designed to assess the retest reliability, tester effects, and operating characteristics as a function of mode of administration (group vs individual) of five test instruments on 2nd and 3rd grade pupils. The instruments included in this experiment were:

- The Gumpgookies Achievement Motivation Test
- The Locus of Control Test
- Two versions (smile-to-frown vs smile-to-neutral) of the Faces Attitude Inventory
- The Coopersmith Self-Esteem Inventory
- The Coloured Progressive Matrices

Among the specific questions addressed by this experiment were:

(a) What are the apparent operating characteristics (mean, variance, reliability, standard error) of each of these instruments?

(b) To what extent is test reliability determined by mode of administration (group vs individual, except the Matrices)?

(c) What are the relative psychometric properties for the two versions of the Faces test?

(d) What are the administrative requirements, caveats, and sources of tester bias for each test with 2nd and 3rd grade "disadvantaged" samples?

(e) What are the apparent unique and common measurement properties among these tests?

Subjects

The subjects were 168 2nd and 3rd grade students from eight classrooms at Peres School in Richmond, California, an economically depressed community. A total of 139 subjects, 70 in Grade 2 and 69 in Grade 3, were pretested. Of these, 120 were given posttests three weeks later. The ETS Locus of Control Test was given to an additional 29 subjects, thirteen in Grade 2 and 16 in Grade 3.

Design

The design was a 2 x 2 x 2 factorial with proportional numbers of subjects in some cells. Two grade levels (2nd and 3rd); two modes of administration (individual and group); and two administrations (test and retest) were employed.

Cell size was determined by the desire for specific kinds of information about test properties and administrative characteristics. Approximately five of every nine subjects were group-tested on both occasions; two out of nine were tested individaally both times; one in nine was tested first individually and then in a group; and one in nine was tested first in a group and then individually (see Figure 2).

Figure 2

TEST PLAN FOR COMPARING EFFECTS OF TESTING MODE--
SECOND TEST OCCASION

		Posttest: Group	Posttest: Individual
Pretest	Group	A 1, 2, 3 B 1, 2	B 3
	Individual	C 1	C 2,3

Note: Letter entries in the table refer to classrooms. Classes have been divided into thirds as indicated by subscripts.

Materials

The Gumpgookies test has 4 practice and 60 test items, with two illustrated alternatives for each item. A running commentary describing activities of little ghost-like figures called Gumpgookies accompanies the test. The commentary is read to the subjects by the Experimenter. In the original and also in the present individual form of the test, the Experimenter records each child's selections on a separate score sheet. In the group form, subjects mark their choice for each item with an X. Test booklets contain one item per page.

The Locus of Control test is made up of an answer sheet on which the Experimenter records the subject's selections for each of the 22 items,

for which there are two alternatives. A child is shown one of twenty-two 8-$\frac{1}{2}$" x 11" drawings of children in school or social situations which is appropriate to his or her sex and race. There are separate sets of pictures for white boys, white girls, Black boys, and Black girls. The original ETS form of the Locus of Control was used in individual testing only.

A thirty-two item version of the Locus of Control test was made by repeating ten items chosen at random from the original 22 and using a single set of 32 pictures to illustrate the 32 test questions. Eight pictures were used from each of four sets in the ETS test (i.e., 8 drawings for Black boys, 8 for white boys, 8 for Black girls, and 8 for white girls). Drawings for repeated items differ with respect to either sex or race (but not both) from those used in the first presentation of these items. Test booklets have one item per page. A drawing covers the top two-thirds of a page; below it are two 2-$\frac{1}{2}$" x 2-$\frac{1}{2}$" black-outlined squares lettered A and B, each enclosing one alternative for the item pictured above. In both individual and group administrations, after the Experimenter has read the item and its alternatives aloud, the subject marks an "X" on the alternative he selects. The booklet form of the Locus of Control was designed for both group and individual administration.

Two versions of the "Faces" Test were prepared, Form A and Form B. Each of the two versions contains the same questions (3 practice and 14 test items) and ask the child to mark the face which shows his feelings about himself and school. The test booklets contain three items per page. Choices for Form A are faces labelled Happy, A Little Happy, and Not Happy; choices for Form B are Happy, Not Happy, and Sad.

The Coopersmith Self-Esteem Inventory was changed in three ways for this experiment. First, eight items referring to parents and home were deleted, leaving a 50-item test. Second, each child marks each item as Like Me or Not Like Me, rather than Like Me and Unlike Me. Third, a practice item was added. Test booklets had ten items on each page.

The only changes made for group administration of the Coloured Progressive Matrices were in the instructions (an adaptation of those used by Tuddenham, 1958, for group administration of this test); and in the score sheet, for which answer spaces were made larger and answer columns separated further than they are on the standard score sheets usually used for the Ravens.

Pencils, red crayons, and test booklets were supplied for each subject.

Procedure

The following procedure was used to assign subjects to the different test conditions. First, each of three classrooms at Grades 2 and 3 was arbitrarily designated as A, B, or C. Room A was tested as an intact classroom on both occasions. Room B was group-tested as an intact classroom in the first testing; for posttesting, two-thirds (2/3) of the students in Room B and one-third (1/3) of the students in Room C were tested as a group. The remaining students in Room B were posttested individually. All students in Room C were given individual pretest two-thirds (2/3) of them were posstested individually.

Two test teams of four members each operated simultaneously. For the pretests, Team A administered group tests to Room A at both grade

levels and individually tested half of the students in Room C. Team B group-tested in Room B and individually tested the other students in Room C. For the posttests, Team A group-tested students in Room A, individually posttested students in Room C whom they had individually pretested, and tested 3rd Graders from Room B specified for individual posttesting. Team B gave group posttests to two-thirds (2/3) of the students in Room B and to one-third (1/3) of those in Room C, and individually tested the remaining students in Room C and 2nd Graders from Room B tested individually.

In the pretests, students in Room A received Form B of the "Faces" Test and students in Room B had Form A. The reverse was true for posttesting. Students in Room C were given either Form A and/or Form B.

All subjects were given the SRI version of the Locus of Control in the pretest; in the posttest, half of the individually tested subjects were given the ETS Locus of Control Test. In addition, at the time of the posttesting, the ETS Locus of Control was individually administered to thirteen 2nd Graders and sixteen 3rd Graders not included in any of the other testing.

Two strategies were used for group testing. Team A group tested the 3rd Grade on consecutive mornings and the 2nd Grade on consecutive afternoons. Team B gave the first three tests in the morning and the last two that afternoon to the same classroom.

The Gumpgookies, Locus of Control, "Faces" Test (both forms), and Coopersmith Self-Esteem Inventory were administered to all subjects given

both pre- and posttests. Only those subjects tested in a group mode were given the Coloured Progressive Matrices. For group testing, the Gumpgookies, Locus of Control, and "Faces" tests were given in that order in a single test session; and the Coopersmith and the Raven's, in that order, were given in a second session. Individual testing of the Gumpgookies, Locus of Control, "Faces," add Coopersmith, in the order named, was completed in a single session.

RESULTS AND DISCUSSION FOR EXPERIMENT 2

PART I: ITEM AND TEST STATISTICS

The Gumpgookies Test

Summary statistics, the mean, standard deviation, reliability (KR 20) estimate, standard error of measurement, and number of cases in each of the administrative modes and the grade levels for the test and retest of the Gumpgookies are presented in Table 4.

The test demonstrates a rather pronounced ceiling effect with an overall mean of 53.36 out of 60 items. The individual means range from a low of 51.63 to 55.05, indicating the effect is present in all subgroups. The measure demonstrates high stability from test to retest (53.29 vs 53.42, respectively) and between grade levels (53.01 for 2nd Grade; 53.74 for 3rd Grade). A small difference is noted in the means of the separate administrative mode groups (54.55 for individually administered measures vs 52.76 for group administered measures). The higher mean for individual measures of affect again suggests social desirability of positive affect may be a significant factor.

It is interesting to note that group administration results in both lower mean performance and greater response variability. The stability of variability estimates between test administrations is also noteworthy. The measure would seem to bave adequate stability.

The estimates of this measure's reliability (.78 to .93) suggest a certain independence from group characteristics, although group administered measures demonstrate the higher estimates.

The percent of positive responses to the Gumpgookies test and retest for each measurement group are consistently high across all both measurement occasions, and all measurement groups, showing a rather strong ceiling effect. Also, responses tend to be quite stable between test occasions.

In summary, the Gumpgookies test presents fairly stable psychometric properties. Its reliability is considered adequate, but the high level of group means and the obvious ceiling effect suggest that the measure may not be appropriate for children of this age range. Social desirability factors may be present in all administrative modes, since the items show little subtlety in masking the socially desirable alternatives. Individually administered forms of the instrument seem to demonstrate special vulnerability to this factor.

Table 4

GUMPGOOKIES TEST: SUMMARY OF OVERALL TEST-RETEST STATISTICS BY GROUP

	Individual			Group		
	2nd Grade	3rd Grade	2nd and 3rd Grades	2nd Grade	3rd Grade	2nd and 3rd Grades
Mean						
Test	54.96	53.57	54.26	51.63	54.20	52.80
Retest	54.71	55.05	54.88	52.73	52.76	52.74
Standard deviation						
Test	5.40	4.68	4.73	8.49	5.94	7.23
Retest	5.18	5.00	4.89	8.26	7.55	7.72
Reliability (KR20)						
Test	.87	.78	.82	.92	.88	.91
Retest	.85	.85	.84	.93	.91	.92
Standard error						
Test	1.96	2.21	2.04	2.39	2.07	2.17
Retest	2.03	1.94	1.93	2.24	2.24	2.17
Number of cases						
Test	23	23	46	49	40	89
Retest	21	20	41	48	41	89

Locus of Control Measures

The summary test statistics of the Locus of Control measures are presented in Table 5. There are two measures reported in this table. The ETS version of the test was administered only in the individual mode, while the SRI version was administered in both modes.

The test means indicate equivalent levels of performance between the testing occasions for both versions of the test. The SRI version demonstrates a slight mode difference, with the individually administered measures being somewhat higher (14.23 vs 13.11). There also seems to be slight grade level effect (14.25 for 3rd grade; 12.70 for 2nd Grade). The ETS version demonstrates the same trend for grade level differences (14.64 for 3rd Grade; 13.50 for 2nd Grade). By comparison, the 3rd and 2nd Grade means for individually administered SRI test forms indicate means of 14.60 and 13.36, respectively. This equivalency of mean performances seems to indicate equivalency of measures.

The reliability measures are quite low for this measure. With the exception of the individually administered 3rd Grade original SRI test, all estimates are fairly stable. In one case, a negative estimate is actually obtained (-.17). With this exception the SRI and ETS forms would again appear equivalent, with perhaps a slight advantage for the SRI version.

Standard error estimates again demonstrate remarkable robustness (a range of 1.61 to 2.00). Even the exceptional test group yields an estimate near the center of this range (1.72). Also, considerable variability in item difficulty occurs both between items and between groups for

any given item, and items tend to demonstrate adequate stability between testing occasions.

Summarizing the Locus of Control data implies that the SRI and ETS versions of the measure are psychometrically equivalent. Although no clear case can be made for the group administration mode over the individual administration mode for the SRI version, the administrative efficiency of the group mode would argue for its adoption, other factors being equal.

Table 5

LOCUS OF CONTROL TEST: SUMMARY OF OVERALL TEST-RETEST STATISTICS BY GROUP

	SRI				ETS	
	Individual		Group		Individual	
	2nd Grade	3rd Grade	2nd Grade	3rd Grade	2nd Grade	3rd Grade
Mean						
Test	13.65	15.00	12.37	13.71	13.62	14.50
Retest	14.25	13.73	12.17	14.48	13.36	14.89
Standard deviation						
Test	2.28	1.59	2.74	2.68	2.33	2.58
Retest	3.05	2.57	2.98	2.41	2.58	2.20
Reliability (KR20)						
Test	.42	-.17	.46	.57	.32	.58
Retest	.67	.56	.57	.50	.50	.47
Standard error						
Test	1.89	1.72	2.00	1.75	1.92	1.66
Retest	1.74	1.70	1.96	1.70	1.82	1.61
Number of cases						
Test	23	24	49	41	13	13
Retest	12	11	47	42	11	9

The "Faces" Attitude Inventory

A summary of the test-retest statistics for the two test forms, two administrative modes, and two grade levels included in the study is presented in Table 6.

Test means ixhibit variability across forms, modes, and grade levels from a low of 8.43 to a high of 12.77 (average = 10.24). The results from the various groups demonstrate considerable score stability from initial test to retest. While no systematic differences are noted between testing occasion or grade level, the two forms and administrative modes show differential results. Averaged over groups, the form means differ by about one and one-half score points, with Form B indicating the higher mean (10.85 vs 9.33). A score difference is also noted between the individual and group administration modes, with individually administered measures indicating a higher level of positive responses (11.33 vs 9.58). It would appear that individually administered measures may elicit a greater number of social acquiescent type responses; that is, in a one-to-one situation, the student may respond in what may be a socially desirable manner rather than how he actually feels concerning items such as "coming to school in the morning." The reliabilities are low enough to suggest the test is only moderate to poor in terms of discriminating individual differences at these grade levels. Variation in standard error remains smal., and comparison of the test-retest standard errors displays the stability of the estimate of the true score confidence interval (i.e., standard error estimates are stable from test to retest, generally change less than 0.2 raw score points). If scores are aggregated to the

classroom level, classrooms whose average positive response scores differ by more than approximately 2 points can be interpreted as being significantly different on this measure.

In summary, the distributional statistics, operating characteristics, and item analyses of the "Faces" test produce results that are neutral to moderately positive. The reliability values (KR 20), while not particularly high, were adequate for group forms and were about the level normally found with similar noncognitive measures. The group administered format of Form B would seem to have the most desirable characteristics. The stability and uniformity of the standard error of measurement estimates suggest the instrument has relatively robust properties and may be useful as a method of detecting group differences.

Table 6

"FACES" TEST: SUMMARY OF OVERALL TEST-RETEST STATISTICS BY GROUP

	Test Form A				Test Form B			
	Individual		Group		Individual		Group	
	2nd Grade	3rd Grade	2nd Grade	3rd Grade	2nd Grade	3rd Grade	2nd Grade	3rd Grade
Mean								
Test	*	11.54	8.88	8.43	12.77	11.00	10.52	9.68
Retest	9.40	10.58	8.93	9.63	12.43	11.25	11.09	9.61
Standard deviation								
Test	*	1.56	2.69	2.27	1.41	1.90	2.74	3.23
Retest	3.02	2.43	4.02	2.99	1.72	2.05	3.57	2.89
Reliability (KR 20)								
Test	*	.38	.62	.59	.58	.49	.78	.80
Retest	.77	.71	.88	.74	.72	.66	.90	.76
Standard error								
Test	*	1.23	1.66	1.46	.91	1.36	1.30	1.45
Retest	1.44	1.31	1.42	1.54	.91	1.19	1.14	1.42
Number of cases								
Test	1	13	24	23	22	11	25	22
Retest	15	12	27	24	7	8	22	23

* Not reported, single case.

The Coopersmith Self-Esteem Inventory

The test-retest statistics for the Coopersmith Self-Esteem Inventory are presented in Table 7, both for the original first form and for an abbreviated form constructed by deleting items deemed inappropriate or ambiguous by a consensus of the testers.

Test means indicate small variability around an overall mean of 29.44 for the full test, indicating no danger of ceiling effects. A slight score decrement is present between test and retest which appears parallel in terms of grade level and test mode. A social desirability factor may have been occurring in the individual testing situation.

The reliability (KR 20) estimates indicate considerable variability, but they are, in general, better than those noted for the "Faces" measure. Group forms again demonstrated great score variability and therefore higher reliability estimates. The group forms at retest demonstrate values of .82 and .92 for 2nd and 3rd Graders, respectively.

Also, the stability of the standard error estimates is again evident. The statistics for the abbreviated form (29 items) indicate that deleting the items changed the properties very little, except for the slight reduction in the reliability estimates. This may be an artifact of the number of items included. If the high and low reliabilities (.60 and .84 are corrected by the Spearman-Brown prophecy formula for a test of 50 items (the original test length), the values become .72 and .90. These values accurately estimate the range of reliability values actually obtained for the 50-item measure.

In summary, the psychometric properties of the Coopersmith Self-Esteem Inventory seem to indicate it is a moderately adequate measure of student affect. While reliability estimates are again discouraging, the stability of the estimates of the error of measurement suggest a robustness independent of group characteristics. It would seem that group administrations are to be preferred. The abbreviated form of this measure would have to be administered alone to make any judgements concerning its adequacy.

Table 7

COOPERSMITH SELF-ESTEEM: SUMMARY OF OVERALL TEST-RETEST STATISTICS BY GROUP

	Original Form				Abbreviated Form	
	Individual		Group		Individual	Group
	2nd Grade	3rd Grade	2nd Grade	3rd Grade	2nd and 3rd Grades	2nd and 3rd Grades
Mean						
Test	28.18	31.12	29.17	31.25	19.96	19.14
Retest	29.82	33.35	27.91	27.69	21.05	18.09
Standard deviation						
Test	6.75	5.14	5.43	6.01	3.56	4.06
Retest	4.64	5.50	7.57	10.28	3.77	5.84
Reliability (KR20)						
Test	.82	.64	.64	.74	.60	.66
Retest	.55	.71	.82	.92	.65	.84
Standard error						
Test	2.84	3.09	3.24	3.05	2.25	2.38
Retest	3.09	2.96	3.21	2.99	2.22	2.31
Number of cases						
Test	22	24	47	44	46	91
Retest	22	20	44	48	42	92

The Raven's Coloured Progressive Matrices Test

The summary statistics for the Matrices test are reported in Table 8.

The test and retest means indicate a small positive increment in correct responses between testing occasions (17.51 on the original test; 18.00 on retest). On this type of measure, one can interpret such change as reflecting learning. Such learning includes test wiseness as a major component, i.e., students are learning to take this type of test. A score difference is also present between grade levels (16.46 for 2nd Grade; 19.15 for 3rd Grade). This difference seems to be the direct result of the higher level of cognitive development of the more mature students. The overall test mean (17.30) indicates that no ceiling effect is operating in this 32-item measure.

The reliability estimates for the measure are high and notably stable. One would expect both of these results for adequate cognitive measures. It is interesting that reliabilities increase from original test to retest with both grade level groups. Standard error estimates show substantial homogeneity and stability.

Inspection of average item difficulties reveals that the overall score difference for the two grade level groups is generally also reflected on an item by item basis, that items in this measure are ordered within item groups (A, AB, A), and the items seem to measure the full difficulty range from very easy to very difficult.

In summary, the Raven's Coloured Progressive Matrices tend to provide consistent evidence of a psychometrically sound cognitive

measurement device. Reliabilities are adequate if not ideal. Item response data indicate that the items are appropriate and reflect the characteristics the test developers desired.

Table 8

COLOURED PROGRESSIVE MATRICES TEST: SUMMARY OF OVERALL TEST-RETEST STATISTICS BY GROUP

	Group 2nd Grade	Group 3rd Grade
Mean		
Test	16.15	19.00
Retest	16.79	19.30
Standard Deviation		
Test	5.79	5.31
Retest	6.41	7.55
Reliability (KR20)		
Test	.84	.84
Retest	.88	.92
Standard Error		
Test	2.33	2.15
Retest	2.24	2.15
Number of Cases		
Test	47	43
Retest	43	43

PART II: COMPARATIVE ANALYSIS OF TESTS

As with Experiment 1 data, scores obtained from these five instruments were organized into intercorrelation matrices to examine for stabilities and for convergent and discriminant validities. Four such matrices are generated: group mode and individual mode data for 2nd grade (Table 9) and for 3rd grade (Table 10) samples. Unfortunately, these correlations do not display the simplex patterns required for straightforward interpretations. It is likely that these sample sizes are much too small, particularly for individual mode data. Moreover, enormous variation in retest reliabilities from 2nd to 3rd grade data are evident in most cases. The few exceptions are the group administered Locus, Coopersmith, and Matrices tests. As such, it appears that the more appropriate test of achievement motivation for 2nd and 3rd grade pupils is the group administered Locus of Control, and the more appropriate test of affect is the group administered Coopersmith. Furthermore, the Matrices test displays appropriate stability (reliabilities range from .71 to .83) as a group administered measure of problem solving.

Table 9

TEST-RETEST CORRELATION MATRIX FOR FIVE MEASURES:
2ND GRADE STUDENTS
(SAMPE SIZE IN PARENTHESES)

Individual Forms	Group Forms: Gumpgookies Test	Gumpgookies Retest	Locus Test	Locus Retest	Faces Test	Faces Retest	Coopersmith Test	Coopersmith Retest	Ravens Test	Ravens Retest
Gumpgookies										
Test		.656* (46)	.159 (42)	.296 (45)	.048 (46)	.167 (46)	.150 (39)	.210 (38)	.104 (34)	.156 (33)
Retest	.152 (20)		.270 (43)	.246 (46)	.183 (47)	.091 (47)	.246 (40)	.361* (39)	.291 (34)	.174 (34)
Locus										
Test	.188 (14)	-.699 (13)		.512* (44)	.344* (44)	-.170 (44)	.198 (36)	.361* (36)	.456* (34)	.372 (34)
Retest	.005 (11)	-.580 (10)	.751* (12)		.004 (47)	-.162 (47)	.181 (39)	.431* (33)	.272 (34)	.054 (34)
Faces										
Test	.143 (21)	.280 (20)	-.091 (15)	-.273 (12)		.203 (48)	.078 (40)	.063 (39)	.287 (34)	.123 (34)
Retest	-.172 (21)	.192 (20)	.198 (15)	.432 (12)	.009 (22)		.046 (40)	.015 (39)	.028 (34)	.179 (34)
Coopersmith										
Test	-.115 (20)	-.196 (19)	-.061 (14)	-.013 (11)	-.301 (21)	.212 (21)		.543* (39)	.266 (32)	.282 (33)
Retest	.161 (21)	.290 (20)	.473 (15)	.492 (12)	-.065 (22)	.368 (22)	-.067 (21)		.368* (32)	.411* (33)
Ravens										
Test										.707* (33)

* $p < .05$.

□ = RELIABILITY (RETEST)

▨ = CONCURRENT VALIDITY

⬚ = DISCRIMINANT VALIDITY

Table 10

TEST-RETEST CORRELATION MATRIX FOR FIVE MEASURES:
3RD GRADE STUDENTS
(SAMPLE SIZE IN PARENTHESES)

Individual Forms	Group Forms: Gumpgookies Test	Gumpgookies Retest	Locus Test	Locus Retest	Faces Test	Faces Retest	Coopersmith Test	Coopersmith Retest	Ravens Test	Ravens Retest
Gumpgookies Test		.311 (36)	.358* (34)	.235 (35)	-.141 (35)	-.022 (35)	.315 (35)	.193 (35)	.191 (27)	.144 (27)
Gumpgookies Retest	.862* (20)		.026 (34)	.179 (38)	.027 (35)	.001 (38)	.158 (35)	.417* (38)	.031 (27)	-.053 (27)
Locus Test	.057 (14)	-.148 (11)		.565* (30)	.093 (34)	.343* (34)	.365* (34)	.340* (34)	.293 (27)	.260 (27)
Locus Retest	-.367 (10)	.051 (10)	.307 (10)		.085 (35)	.296 (38)	.255 (35)	.328* (38)	.346 (27)	.192 (27)
Faces Test	.027 (23)	-.337 (20)	.557* (14)	.170 (10)		.573** (41)	.087 (37)	.185 (37)	-.228 (27)	-.384* (27)
Faces Retest	-.095 (10)	-.275 (20)	.502 (11)	.346 (10)	.374 (20)		.287 (37)	.304 (40)	.028 (27)	-.06 (27)
Coopersmith Test	-.126 (22)	-.305 (19)	-.048 (13)	.329 (9)	.235 (22)	.527* (19)		.652** (41)	.381* (31)	.367* (30)
Coopersmith Retest	.325 (19)	.021 (19)	.312 (10)	-.459 (9)	.443* (19)	.445* (19)	.020 (19)		.378* (31)	.368* (30)
Ravens Test										.830* (30)

* $p < .05$.
** $p < .01$.

= RELIABILITY (RETEST)

= CONCURRENT VALIDITY

= DISCRIMINANT VALIDITY

PART III: COMPARISON OF ADMINISTRATIVE PROBLEMS AMONG THE FIVE MEASURES

To provide some basis for evaluating the administrative aspects of the various measures, test administrators were asked to critique each of the measures and to note administrative difficulties they encountered during the testing. The following paragraphs summarize their observations.

Gumpgookies

The Gumpgookies measure was unanimously noted as being too long and testers said children quickly became bored. Student boredom seemed to be both a function of test length and inappropriateness of the story for children of this age level. Content would seem to be more appropriate for a younger population. It was also noted by all of the testers that children seemed to be quite aware of the socially accepted answer. The following quote from a tester summary (group testing) seems indicative of the consensus of testers:

> "...because the test was so long, because of the obvious (socially accepted) answers, and because of the marking procedures, discipline became a problem. Many children went ahead of the experimenter and consequently lost their places. Other children would share their responses with the entire class."

Locus of Control

Tester comments indicated that the Locus of Control measure, particularly the SRI group form, prescribed a number of administrative difficulties. The majority of the testers found the measure somewhat long and tedious to administer because of the repetitive item format. All testers again noted social desirability as a rather significant potential contaminant.

The physical structure of the group-administered SRI version of the measure prompted several negative comments. In particular, it was noted that the physical size of the booklet overwhelmed the children and that the large response boxes prompted children to look at one another's answers. The stimulus pictures were not of sufficient quality to allow any ethnic differences in the characters to be detected.

Testers also noted that certain items (3, 5, 7, 9 and 22) may be interpreted as being classroom specific and situationally dependent. Therefore the locus is a priori external.

While the individual forms (both SRI and ETS) seemed administratively more manageable, the social acceptability factor seemed to be strongly evident in individual testing.

"Faces"

Testers found the "Faces" measure easy to administer, code, and score. Group administration seemed to work well. The major criticism of the measure was that response bias might determine the choices students make. In particular, testers found children marking only the happy face, sometimes even before the tester had read the question. This observation is supported by the high number of positive choices noted in the data. While testers preferred Form A (no frown) over Form B, they generally questioned the usefulness of the measure in terms of the meaning and range of response alternatives.

Coopersmith Self-Esteem Inventory

The Coopersmith instrument was unique in having made a generally positive impression upon the testers. It was, in fact, felt by all of the testers to be the best noncognitive measure of the four affective tests given. Ease of administration in both individual and group modes was cited as one positive attribute. Minor format changes were suggested for revising the measure, but format was not cited as being a strongly negative attribute.

Three noteworthy aspects of the measure were commented on by most of the testers. First, testers felt that the children were more personally involved in responding to the items than with the other noncognitive measures. Typical tester comments include:

> "...the children appeared to think very carefully about the items and mark the ones that applied to them..."

> "The children seemed to attend to the task due to its personal content."

Second, testers felt that social desirability and response bias did not seem to play as great a role in this measure as with the others. There seemed to be fewer instances of all positive responses. This result is probably due, in part, to the increased personal involvement noted above, and it was noted particularly in the group mode. In individual testing, the students seemed more conscious of the tester's expectations.

The third aspect generally commented on by testers was the wording of certain items. They noted the use of double negatives and the difficulty of the vocabulary in many items. It would seem that to make the measure wholly appropriate to the age level of the children in this study, some item revision should be undertaken. In the analysis of test statistics reported earlier, the items deemed troublesome by the testers were deleted and the data reanalyzed (abbreviated form). It should be noted again that this did not seem to adversely affect the overall psychometric properties of the measure. It seems reasonable to believe that the measure could be quite useful with the item changes notes.

Raven's Coloured Progressive Matrices

Testers noted the administrative ease of the Raven's Matrices as a very positive factor of this measure. Furthermore, they noted that the students seemed to enjoy the challenge of the measure. The major negative comment was that allowing children to pace themselves resulted in many children competing to be first. Time seemed more important to some children than accuracy. It was suggested that the format could be changed to pace the students. For example, one could allow them to work on only a limited number of items in each of several controlled time periods.

Answer sheets also drew some criticism and suggestions for directly marked test sheets were numerous. The cost involved in using colored stimuli was also questioned, and the use of black and white expendable forms was suggested.

The testers generally felt that the items were appropriate but that the last two or three items in each set were perhaps too difficult and might be eliminated. It was felt the removal of these items might lessen test anxiety.

SUMMARY AND RECOMMENDATIONS

Summary

In general, all of the measures had specific drawbacks, which are noted in the comments made by the testers observing students' reactions to the measures. Administrative problems were often cited, and the four noncognitive measures were considered especially vulnerable to the effects of social desirability. It was felt that the Coopersmith Self-Esteem Inventory, whatever its other weaknesses, was notable in minimizing these effects, particularly in the group administration.

Tester comments and statistical analysis alike indicate that group administration is preferable, not only in terms of the psychometric properties that the measures display but also in terms of efficiency of testing; that is, on the basis of the data presented here, group tests are more efficient and tend to have better psychometric properties as well. This is certainly due in large measure to the anonymity which the group process provides for noncognitive measures.

The Gumpgookies test is clearly inappropriate for the more mature group of children. This is evident both from the extremely high level of positive choices noted and from the tester comments. The "Faces" and, to a lesser extent, the Locus of Control measures also appear to exhibit the social desirability and response bias problems. The Coopersmith Inventory generally yielded more favorable results, both in terms of a statistical analysis of its psychometric properties and in terms of the tester comments. Minor item revisions should render this measure quite useful. The Matrices measure already has been well researched and documented, and the present study further supports conclusions

regarding its psychometric properties and its appropriateness as a group administered test of problem solving behaviors.

Recommendations

Based on the statistical and anecdotal evidence obtained from this second experiment, we offer the following recommendations regarding the adoption of these instruments for a testing program at the 2nd and 3rd grade levels.

Achievement Motivation--The Locus of Control measure appears reasonably suited to administration in a group mode to 2nd and 3rd grade children. Care should be exercised to counterbalance item formats so that position effects are minimized. Our data do not support the argument that separate versions are needed for ethnic and gender differences. The caveat of social compliance intrusions is less relevant to this instrument, since to some extent the construct (achievement motivation) is a function of perceived social/cultural expectancies.

Our data indicate that the Gumpgookies test is inappropriate for administration at the 2nd and 3rd grades. We believe this test is more appropriate to a younger population (preschool or kindergarten).

Affect--Our results show the group administered Coopersmith to be a reasonably appropriate test of attitudes toward self, school, peers, and home for 2nd and 3rd grade pupils. We feel certain ambiguous or linguistically complex (i.e., double negative) items can be revised to reduce measurement error and further enhance interpretability of

resultant scores. We also feel that a shortened version (29 items) of this test would be adequate.

The Faces test, on the other hand, does not appear well suited to children at this age level. But since the Faces test exhibited generally favorable results in Experiment 1, it appears that it is appropriate for use with younger children (again preschool or kindergarten) under an individual mode of administration.

Problem Solving--As noted earlier, the Matrices test appears well suited as a group administered measure of problem solving at these grade levels. Furthermore, deletion of several items from each section would seem to (a) shorten administration time without substantially altering the reliability and validity of scores (the items are so difficult that virtually none is passed, hence no measurement), and (b) eliminate the competing interpretation of scores as reflecting general intelligence (the test becomes altered, therefore IQ conversion tables would no longer apply).

REFERENCES

Adkins, D.C. and Ballif, B.L. Motivation to achieve in school. Final Report to the Office of Economic Opportunity, Contracts B89-4576 and 4121. January 1970.

Adkins, D.C. and Ballif, B.L. A new approach to response sets in analysis of a test of motivation to achieve. Educational and Psychological Measurement, 1972, 32, 559-577.

Brown, B.R. The assessment of self concept among four-year old Negro and White children: a comparative study using the Brown-IDS Self Concept Referents Test. Paper presented at the Eastern Psychological Association Meeting, N.Y.C., April, 1966.

Coopersmith, S. The antecedents of self-esteem. San Francisco: W.H. Freeman, 1967.

Cronbach, L.J., The Essentials of Psychological Testing, Harper & Row, 1970, p. 161 f.

Emrick, J.A., Sorensen, P.H,, and Stearns, M.S. Interim evaluation of the national Follow Through program 1969-1971. Menlo Park, California: Stanford Research Institute, 1973.

Hertzig, M.E., Birch, H.G., Thomas, A., and Mendez, O.A. Class and ethnic differences in the responsiveness of preschool children to cognitive demands. Monographs of the Society for Research in Child Development, 1968, Ser. No. 117, Vol. 33, No. 1.

Kirk, S.A., McCarthy, J.J., and Kirk, W.D. Examiner's Manual: Illinois Test of Psycholinguistic Abilities (Revised Edition). Urbana: University of Illinois Press, 1969.

Maehr, M.L. and Sjogren, D.D. Atkinson's theory of achievement motivation: First step toward a theory of academic motivation. Review of Educational Research 1971, 41 (2), 143-161.

Paraskevopoulos, J.N. and Kirk, S.A. The Development and Psychometric Characteristics of the Revised Illinois Test of Psycholinguistic Abilities. Urbana: University of Illinois Press, 1969.

Stanford Research Institute. General instructions and manual of procedures for Follow Through testing. Menlo Park, California: Author, Fall 1969.

Stanford Research Institute. Issues of non-cognitive measurement. Menlo Park, California: Author, December 1970.

Trowbridge, N. Self concept and scoio-economic status in elementary school children. American Educational Research Journal. 1972, 9 (4), 525-537.